УДК 621.396.4
DOI 10.18799/29495407/2023/2/27

ОЦЕНКА ИСПОЛЬЗОВАНИЯ ИНСТРУМЕНТОВ БИБЛИОТЕКИ SPACY И DEEPPAVLOV ДЛЯ ЗАДАЧИ ИЗВЛЕЧЕНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ ИЗ ОПИСАНИЙ РЕЗУЛЬТАТОВ ОСМОТРОВ ПАЦИЕНТОВ С COVID-19

Соколовский Дмитрий Евгеньевич[1],
desokolovskii@gmail.com

Некрасов Владимир Николаевич[2],
nekrassov@yandex.ru

Землянский Сергей Александрович[3],
qoelky@gmail.com

Аксёнов Сергей Владимирович[1],
axyonov@tpu.ru

[1] Национальный исследовательский Томский политехнический университет,
Россия, 634050, г. Томск, пр. Ленина, 30.

[2] Центр клинико-лабораторной диагностики Военно-медицинской академии им С.М. Кирова,
Россия, 194044, г. Санкт-Петербург, ул. Академика Лебедева, 6, лит. Ж.

[3] Национальый исследовательский Томский государственный университет,
Россия, 634050, г. Томск, пр. Ленина, 36.

Актуальность. *Определяется необходимостью выделения значимых признаков из электронных медицинских записей для автоматизации оценки состояния больных.* ***Цель.*** *Оценка возможности выявления именованных сущностей в электронных описаниях осмотров пациентов с COVID-19 с помощью модели BERT из библиотек SpaCy и DeepPavlov.* ***Методы.*** *Глубокое обучение, статистические методы.* ***Результаты и выводы.*** *Выполнено исследование настройки нейросетевых моделей BERT из библиотек SpaCy и DeepPavlov для аннотирования документов «Осмотр пациентов лечащим врачом» с целью выделения следующих предикторов оценки состояния пациентов: температура, артериальное давление, частота дыхательных движений, частота сердечных сокращений и сатурация. Настройка и оценка эффективности архитектур производилась на основе разметки 340 обезличенных электронных медицинских записей пациентов, болевших COVID-19, полученных с помощью сервиса SibMED Data Clinical Repository. Показано, что настройка моделей на количестве около 150 размеченных документов позволяет определять указанные предикторы в таких текстах с точностью (Precision) 85–98 % и с полнотой (Recall) 77–98 % в зависимости от предиктора. Метрики качества работы архитектур из выбранных библиотек различались незначительно. Отмечено, что итеративное расширение обучающей выборки в результате эксплуатации моделей с последующей донастройкой приводит к повышению результативности моделей.*

Ключевые слова: *Глубокое обучение, извлечение именованных сущностей, BERT, SpaCy, DeepPavlov.*

Введение

Одной из важных причин роста популярности технологий машинного обучения является накопление больших массивов неструктурированных данных во многих аспектах деятельности человека. Многие компании (включая те, бизнес которых традиционно не относят к сфере информационных технологий) вкладывают существенные ресурсы, связанные с анализом данных, получившихся и получающихся в результате их бизнес-процессов, для повышения своей эффективности, мониторинга качества работ и выделения новых зависимостей в данных, которые могут привести к новым интересным результатам. Здравоохранение и медицина относятся к таким отраслям, которые генерируют значительный объём разнородных данных (результаты многочисленных анализов в бумажном и электронном виде, ЭКГ, КТ, дневники самонаблюдений, осмотры специалистом и т. д.) [1]. Анализ таких разнородных данных зачастую весьма затруднён, даже в случае их наличия в электронном виде. Существует большое количество проблем, связанных со сложностью анализа подобных документов, включая разные форматы хранения, варианты кодирования информации, сложность представления и др. К этому добавляется человеческий фактор, заключающийся в текстовой фиксации значимых предикторов [2]. Например, в документе «Осмотр пациента лечащим врачом», специалист может отметить, что «температура была 38.5 градусов» или «тем-ра 38,5» или « t=38.5 С». Из-за значительной нагрузки на лечащих врачей многие из них при заполнении медицинской документации могут использовать шаблоны документов, которые необходимо подкорректировать под конкретного пациента, сокращать наименования, пропускать буквы в словах или допускать грамматические ошибки. При создании интеллектуальных приложений, выполняющих анализ разнообразных признаков пациентов, необходимо сначала организовать выделение этих предикторов с помощью программных средств.

Структуризация таких данных возможна через процесс распознавания именованных сущностей (NER – Named Entity Recognition) [3]. Классически к задачам

NER относят выделение специфических объектов, таких как, например, имена людей (Мария, Сабина Сафина), организаций (ТПУ, ФНС, Яндекс), местоположения (Москва, Нью-Дели, Бразилия), денежные значения (1000 долларов США, пять тысяч руб.) названия специфических продуктов компаний [4]. Однако этот процесс можно применить для выбранной исследователем предметной области для получения сущностей, являющихся специфическими для неё.

NER также используется для анализа медико-биологических документов с целью извлечения значимых слов, таких как наименование лекарств, жалобы, результаты объективного осмотра, значения лабораторных показателей т. д. [5–7].

На рис. 1 показан фрагмент текста, из которого необходимо выделить определенные сущности.

Фавипавир применяется в системах лечения в амбулаторных условиях при SARS-COV2. Перед приёмом данного препарата было классифицировано тяжёлое течение заболевания показатель насыщения крови кислородом (SpO2) у Иванова Андрея Михайловича был 92%. Температура тела держалась на уровне 38,5С. Наблюдались изменения в лёгких при КТ (рентгенографии), типичные для вирусного поражения.

Рис. 1. *Фрагмент анализируемого текста. Именованные сущности подчеркнуты*

Fig. 1. *Fragment of the analyzed text. Named entities are underlined*

Результат работы алгоритма выделения NER представляет собой сопоставление частей текстовых последовательностей с релевантными классами, нужными исследователю. Для фрагмента, приведенного на рис. 1, желаемый результат работы алгоритма NER представлен в табл. 1.

Таблица 1. *Классы сущностей и найденные именованные сущности для текста из фрагмента на рис. 1.*

Table 1. *Entity classes and found named entities for the text from the fragment in Fig. 1.*

Класс сущности Entity class	Найденная именованная сущность Found named entity
Препарат/Drug	Фавипавир/Favipavir
Диагноз/Diagnosis	SARS-COV2
Тяжесть состояния Severity of the condition	Тяжёлое течение/Heavy current
ФИО/Full name	Иванова Андрея Михайловича Ivanov Andrey Mikhailovich
Единица измерения/Unit	%, С
Показатель «Температура» Indicator "Temperature"	Температура тела/Body temperature
Значение температуры/Temperature value	38,5
Показатель «SpO2» "SpO2" indicator	SpO2
Значение SpO2/SpO2 value	92
Комментарии/Comments	изменения в лёгких при КТ (рентгенографии), типичные для вирусного поражения changes in the lungs on CT scan (x-ray), typical of a viral infection

Средства технологий NER включают в себя разные подходы к извлечению фрагментов текстов, начиная с классического подхода – использование регулярных выражений, заканчивая мощными средствами нейронных сетей [8]. На практике применение архитектур глубокого обучения показывает наиболее высокие результаты [9–11].

В предлагаемой работе рассматривается использование инструментов глубокого обучения для извлечения наиболее общих предикторов пациента, находящихся в документе «Осмотр пациента лечащим врачом» на русском языке.

Объекты и методика исследования

В настоящее время среди наиболее используемых средств для извлечения именованных сущностей и обработки текстов на русском языке выделяются следующие библиотеки: DeepPavlov BERT NER, slovnet BERT NER, Pullenti, Stanza и SpaCy [4]. По результатам оценки указанных инструментов на наборе данных factru, в который входили тексты на различные темы [12], и наборе данных ne5, в который входили тексты, содержащие разметку имен людей (Per), организации (Org) [13], с помощью библиотеки «Naeval» [14] получены значения метрики F1-score по токенам (минимальным фрагментам анализируемого текста), представленные на рис. 2.

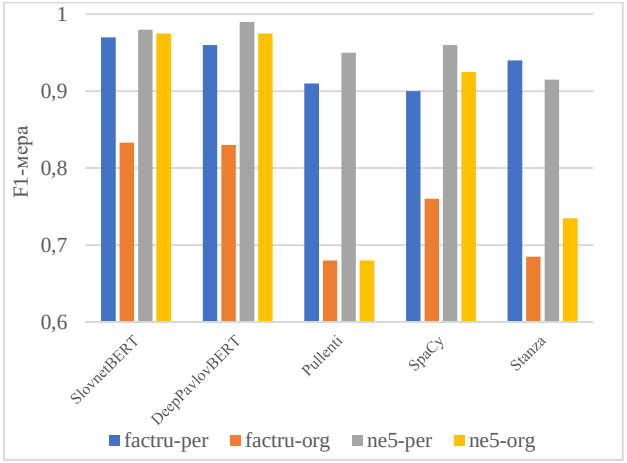

Рис. 2. *Сравнение библиотек для извлечения именованных сущностей на наборах factru и ne5 для тегов личность (per) и организация (org)*

Fig. 2. *Comparison of libraries for extracting named entities on factru and ne5 sets for person (per) and organization (org) tags*

В качестве средства, использованного для нахождения именованных сущностей, в настоящей работе были выбраны библиотеки SpaCy и DeepPavlov, т. к., согласно произведенной оценке, библиотека SpaCy среди средств, разработанных не для русского языка, предоставляет возможность нахождения тегов «Персона» и «Организация», сравнимую с русскоязычными инструментами SlovnetBERT и DeepPavlovBERT. Библиотека DeepPavlov предложена как инструмент, реализованный для русского языка, показывающий высокие прогностические результаты, а также позволяющий использовать латиницу в анализируемых текстах. Решение использовать SpaCy также объяснялось наличием хорошего интерфейса по настройке нейросетевой модели анализа текстов на новых корпусах размеченных текстов.

BERT модели, входящие в состав архитектур, обучены на корпусах общих текстов. Предобученные архитектуры для задач анализа специализированных текстов не могут подходить, т. к. в составе последних выделяемые токены обладают контекстами, отличающимися от контекстов токенов для общих текстов (страниц wikipedia, лент новостей, произведений литературной классики). Архитектура библиотеки SpaCy обучена на корпусах текстов федеральных новостных порталов – ru_core_news_lg [15], для DeepPavlov взята модель ner_ontonotes_bert_mult [16].

По причине для выявления зависимостей между значимыми фрагментами и локализации требуемых сущностей требуется выполнить дополнительную настройку модели на корпусах текстов медицинских осмотров. Так как результатом работы модели должны выступать выделенные фрагменты текста (слова, цифры, символы и словосочетания), анализируемый набор данных необходимо разметить, т. е. выделить области текста, связав их с меткой NER. На рис. 3 показан пример выполненной разметки фрагмента, используемый для настройки нейросетевой архитектуры BERT, который используется в SpaCy.

{"classes": ["Темп", "ЗначениеТемп", "SpO2", "ЗначениеSpO2"], "annotations": [["Объективный осмотр: Рост 166 см, вес 79 кг. Состояние больной средней тяжести. В ясном сознании, вялая. \r\nНа вопросы отвечает правильно. Ориентирована во времени, пространстве. Очаговых и менингеальных \r\nзнаков нет. Гиперстенического телосложения. Кожные покровы обычной окраски, влажные, горячие, тургор снижен. \r\nТемпература тела 38,1 С. Зев гиперемирован, гиперемия дужек, задней стенки глотки, \r\nминдалины не гипертрофированы, налетов нет. … Селезенка не \r\nпальпируется. Мочеиспускание свободное, безболезненное. Симптом поколачивания отрицательный с двух \r\nсторон. Стул 1 раз в сутки.\r\nSpO2= 93% на атмосферном воздухе.", {"entities": [[315, 331, "Темп"], [332, 339, "ЗначениеТемп"], [934, 939, "SpO2"], [940, 943, "ЗначениеSpO2"]]}]]}

Рис. 3. *Пример разметки фрагмента осмотра в формате json, использованный для обучения модели SpaCy*

Fig. 3. *Example of inspection fragment markup in json format used for SpaCy model training*

В блоке «classes» указаны метки NER тех сущностей, которые должны быть извлечены из текстов осмотров. В блоке «annotations» представлена область текста, из которой извлекаются NER, указанные далее в блоке «entities». Связывание метки NER с областью текста производится путем указания позиции начального символа и позиции заключительного символа, входящих в состав именованной сущности для текста в «annotations».

На рис. 4. представлена размеченная часть текста из рис. 3, но предлагаемая для настройки нейросетевой модели BERT от DeepPavlov.

Текст для настройки модели DeepPavlov проаннотирован следующим образом. В конце фрагмента текста, являющегося строкой и не являющегося релевантным объектом, устанавливается тег «O». Именованная сущность, входящая в документ, записывается с новой строки и после неё указывается её тег. Например «Temp» – тег температуры, «SpO2_Value» – тег значения показателя SpO2, и т. д.

В качестве токенизатора для моделей выбрана процедура tok2vec, т. к. русский язык предоставляет много словоформ, а также в тексте значимые слова часто сокращаются, возможны орфографические ошибки и пропуски в словах, что объясняется большой нагрузкой специалистов, заполняющих документацию.

Объективный осмотр: Рост 166 см, вес 79 кг. O
Состояние больной средней тяжести. В ясном сознании, вялая. \r\n O
На вопросы отвечает правильно. Ориентирована во времени, пространстве. Очаговых и менингеальных \r\n O
знаков нет. Гиперстенического телосложения. Кожные покровы обычной окраски, влажные, горячие, тургор снижен. \r\n O
Температура тела Temp
38,1 C Temp_Value
. O
Зев гиперемирован, гиперемия дужек, задней стенки глотки, \r\n O
миндалины не гипертрофированы, налетов нет. … Селезенка не \r\n O
пальпируется. Мочеиспускание свободное, безболезненное. Симптом поколачивания отрицательный с двух \r\n O
сторон. O
SpO2 SPO2
= O
93% SpO2_Value
на атмосферном воздухе O

Рис. 4. *Пример разметки фрагмента осмотра в формате txt, использованный для обучения модели DeepPavlov*

Fig. 4. *Example of inspection fragment markup in txt format used for DeepPavlov model training*

В табл. 2 приведены сущности, извлекаемые из документов в предлагаемом исследовании, примеры сущностей, и теги, использованные при аннотировании документов.

Для обучения подготовлены 340 фрагментов документов «Осмотр пациента лечащим врачом», содержащие записи о пациентах, перенесших COVID-19, и полученных с помощью сервиса SibMED Data Clinical Repository [17]. Этот набор текстов разделен на непересекающихся три поднабора – Set_1 (из 100 текстов), Set_2 (из 100 текстов) и Set_3 (из 140 текстов). Тексты в наборе Set_1 были проаннотированы с указанием тегов из табл. 2 для формирования выборки для настройки нейросетевых архитектур BERT из библиотек SpaCy и DeepPavlov, согласно процедурам разметки, указанным выше. Далее этот аннотированный набор был разделен в отношении 80 к 20 для обучающей и валидационной части. В качестве функции потерь, минимизируемой оптимизационной процедурой, выбрана категориальная кросс-энтропия:

$$J_{catXENT} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} d_j^{(i)} \log(y_j^{(i)}),$$

где N – общее количество сущностей; C – количество классов сущностей; $d_j^{(i)}$ – бинарный индикатор, указывающий на принадлежность примера к классу j (0 – если сущность (i) действительно принадлежит классу j или 0 в противоположном случае); $y_j^{(i)}$ – выход нейросетевой модели, указывающий на вероятность принадлежности сущности (i) к классу j.

Для настройки весовых коэффициентов как моделей BERT, так и выходного слоя архитектуры был использован метод адаптивной оценки момента Adam [18], основанный на оценке экспоненциального скользящего среднего градиентов и квадратов градиентов функции потерь для корректировки весовых коэффициентов.

Таблица 2. *Виды именованных сущностей, извлекаемых моделями SpaCy и DeepPavlov, примеры и имена тегов, использованных при разметке текста для настройки*

Table 2. *Types of named entities extracted by SpaCy and DeepPavlov models, examples and names of tags used when marking up text for training*

Извлекаемая сущность Entity to retrieve	Примеры Examples	Наименование тега для модели SpaCy/DeepPavlov Tag name for the model SpaCy/DeepPavlov
Температура тела Body temperature	Темп., т-ра,темпер. Temp., t, temper	«Темп»/«Temp»
Значение температуры Temperature value	37.9 С, сорок град. 37.9 C, forty degree	«ЗначениеТемп» «Temp_Value»
Артериальное давление – АД Blood pressure	АД, арт.дав. BP, Blood press.	«АД»/«Blood_Press»
Значение показателя АД Blood pressure value	145/85 мм.рт.ст., 120/80 мм рт ст 145/85 mill. Of merc. 120/80 mill merc	«ЗначениеАД» «Blood_Press_Value»
Частота дыхательных движений – ЧДД Respiratory frequency	ЧДД, част.дд RespR, Resp. rate	«ЧДД»/«Resp_Rate»
Значение показателя ЧДД Respiratory frequency indicator value	22 в минуту, двадцать в мин. 22 per minute, twenty per min.	«ЗначениеЧДД»/«Resp_Rate_Value»
Частота сердечных сокращений – ЧСС Heart rate	ЧСС, част сс HeartR, heart rate	«ЧСС»/«Heart_Rate»
Значение показателя ЧСС Heart rate indicator value	78 уд./мин., восемьдесят ударов 78 beats/min, eighty beats	«ЗначениеЧСС»/«Heart_Rate_Value»
Насыщение крови кислородом Blood oxygen saturation	SpO2, SPO 2	«SpO2»/«SpO2»
Значение сатурации артериальной крови кислородом Value of arterial blood oxygen saturation	93 %, 99 %	«ЗначениеSpO2»/«SpO2_Value»

Обучение моделей производилось с помощью видеоускорителя Nvidia Tesla T4 с микроархитектурой Turing, оснащенной тензорными ядрами для ускорения обучения глубоких архитектур.

Дополнительно в процессе настройки анализировалась метрика верность ($Accuracy$):

$$Accuracy = \frac{CC}{AC},$$

где CC – количество именованных сущностей, корректно распознанных моделью и AC – общее количество сущностей, размеченных в выборке. Время обучения архитектуры BERT из библиотеки DeepPavlov на наборе Set_1 для получения весовых коэффициентов, при которой величина $Accuracy$, достигнув 83,2 %, перестала расти на валидационной выборке, оставило около 26 минут. При использовании аналогичного оборудования модель BERT из SpaCy настроилась за 18 минут при $Accuracy$ 78,6 % для той же валидационной части. Отметим, что при достижении этих величин верности в двух моделях наблюдалось значение этой метрики близкое к 100 % на обучающей части, т. е. модель переобучалась.

Результаты исследования и их обсуждение

Тестирование обученной модели производилось на наборе Set_2. Значение метрики $Accuracy$ на этом наборе составило 67,5 %. Назначение набора Set_2 состояло также в выявлении тех сущностей, которые распознаются неверно из-за особенностей анализируемых документов (разные специалисты, заполняющие документы, ошибки (в том числе и орфографические)), которые не присутствовали в наборе Set_1. После оценки качества работы архитектур на наборе Set_2 и получения разметки произведена ручная оценка нахождения сущностей в аннотированном моделью наборе Set_2. Метки не найденных и ошибочно классифицированных сущностей были откорректированы, и исправленный набор был объединен с Set_1. Подобная процедура выполнялась с целью пополнения обучающих примеров новыми образцами, существенно отличающимися от тех, которые отсутствовали в первом наборе Set_1, также такая процедура облегчает аннотирование, т. к. работа по выделению значимых сущностей частично ложится на нейронную сеть.

На рис. 5 показан пример работы модели из библиотеки SpaCy на фрагменте тестового документа в среде Jupyter Notebook. Средства библиотеки SpaCy позволяют визуализировать аннотацию выделением NER с указанием тега. Красным цветом выделены теги названий предикторов, а жёлтым – величины показателей с единицей измерения.

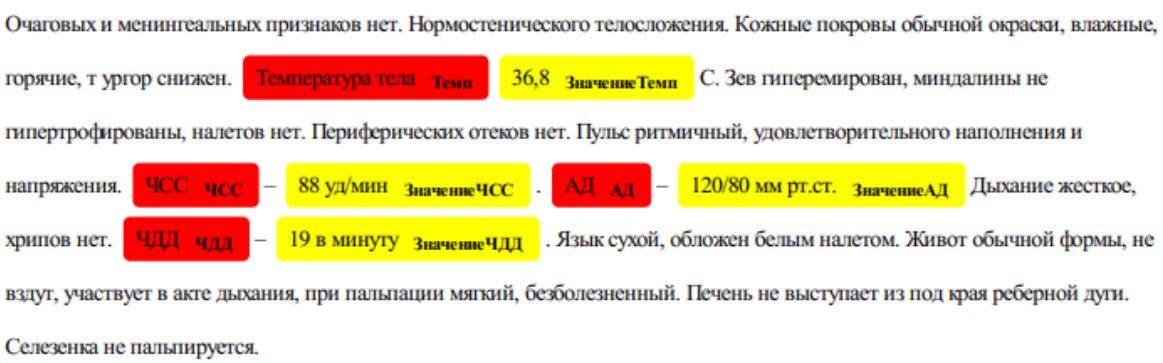

Рис. 5. *Результат аннотирования моделью SpaCy фрагмента документа «Осмотр пациента лечащим врачом»*

Fig. 5. *Result of annotation by the SpaCy model of a fragment of the document «Examination of a patient by the attending physician»*

Из рис. 5 видно, что сущности с тегами «ЗначениеЧСС», «ЗначениеАД» и «ЗначениеЧДД», найденные моделью, состоят из собственно значения и единицы измерения, а сущность с тегом «ЗначениеТемп» включает только величину показателя. При корректировке единица измерения «С» была добавлена в набор. Ошибка определения могла быть вызвана тем, что символ «С» (латиница) и символ «С» (кириллица) располагаются на единой клавише клавиатуры, а также наличием дополнительного пробела между значением и единицей измерения.

При повторной настройке объединенным аннотированным набором Set_1+Set_2 величина *Accuracy* на валидационной выборке из этого объединённого набора достигла 94,3 % для модели из Spacy и 96,2 % для DeepPavlov.

Окончательная оценка производительности выполнена на текстах, до этого не исследованных моделями и входящих в Set_3. Параметр *Accuracy* составил 92,1 % для модели SpaCy и 93,3 % для DeepPavlov. Кроме того, выполнены расчеты метрик точности (Precision) и полноты (Recall) для каждой именованной сущности, чтобы оценить те из них, которые определяются лучше и хуже всего:

$$Precision = \frac{TP}{TP+FP},$$
$$Recall = \frac{TP}{TP+FN},$$

TP – количество правильно классифицированных объектов релевантного класса; FP – количество объектов других классов, ошибочно отмеченных как релевантные; FN – количество ошибочно классифицированных объектов релевантного класса сущности. Результаты анализа приведены в табл. 3.

Таблица 3. *Метрики оценки качества работы моделей SpaCy и DeepPavlov для задачи распознавания именованных сущностей*

Table 3. *Metrics for assessing the performance of SpaCy and DeepPavlov models for the named entity recognition task*

Именованная сущность Named Entity	SpaCy		DeepPavlov	
	Precision	Recall	Precision	Recall
	%			
Температура тела/Body temperature	91,6	88,9	92,3	90,9
Значение температуры Temperature value	85,4	77,0	86,4	81,6
Артериальное давление – АД Blood pressure	98,3	98,6	99,2	97
Значение показателя АД Blood pressure value	90,2	89,3	93,4	93,5
Частота дыхательных движений – ЧДД Respiratory frequency	92,5	95,4	93,1	95,6
Значение показателя ЧДД Respiratory frequency indicator value	91,2	94,7	93,4	94,7
Частота сердечных сокращений – ЧСС Heart rate	93,9	93,6	98,2	94
Значение показателя ЧСС Heart rate indicator value	93,3	92,4	98,0	91,7
Насыщение крови кислородом Blood oxygen saturation	100			
Значение сатурации артериальной крови кислородом Value of arterial blood oxygen saturation				

Как видно из таблицы, метрики моделей SpaCy и DeepPavlov, полученные на наборе Set_3, сравнимы друг с другом, при этом как показатель Precision, так и показатель Recall немного выше для DeepPavlov, что объясняется, скорее, предварительной настройкой, полученной при исследовании текстов общей тематики на русском языке.

После настройки моделей на наборе Set_1 было зафиксировано, что значительная часть ошибок была связана с ошибочной классификацией сущностей температура и значение показателя температура (например, 19 % первых и 38,5 % вторых не выявлено моделью от DeepPavlov в наборе Set_2). Корректировка аннотаций в Set_2 и дополнительная настройка моделей на исправленных примерах позволили повысить качество распознавания. Дополнительно при тестировании моделей выявлено, что они размечали также те части текста, связанные с температурой, которые не требовалось находить (при подготовке данных требовалось найти текущую температуру пациента в блоке «Объективный осмотр» документа «Осмотр пациента лечащим врачом»). Сущности, связанные с температурой, не были размечены в блоке анамнеза заболеваний и жалоб, но модель их обнаружила в этих разделах. С одной стороны, это уменьшило точность модели, но, с другой стороны, это показало, что можно выделять дополнительные признаки из других разделов нейросетевыми моделями.

В четырех документах для модели SpaCy и в трех документах для модели DeepPavlov из набора Set_2 значение показателей ЧСС и ЧДД были перепутаны моделью (т. е. показатель предиктора ЧСС отмечался как значение ЧДД и наоборот). После донастройки на скорректированной разметке, при работе моделей на Set_3, этого не наблюдалось.

Следует отметить, что название и значение показателя сатурации артериальной крови кислородом всегда локализовалось корректно, что, скорее, объясняется расположением этих сущностей в тексте и некой стандартизацией написания этой сущности.

Заключение

По результатам экспериментов, связанных с выявлением именованных сущностей, находящихся в историях болезней, таких как: показатели температуры, частоты дыхательных движений, артериальное давление, частота сердечных сокращений и сатурации, с помощью BERT моделей из библиотек SpaCy и DeepPavlov выявлено, что при разметке около 150 документов «Осмотр пациента лечащим врачом» можно выявить немногим более 90 % указанных именованных сущностей, как моделью из библиотеки SpaCy, так и DeepPavlov. Расширение набора на новые примеры, которые были ошибочно классифицированы, позволило увеличить точность работы моделей. Такой путь, заключающийся в итеративном добавлении в выборку новых неисследованных примеров может использоваться не только для сущностей – предикторов состояния пациента, но и для других объектов. Также выявлено, что сущности температуры и их значения, даже при их разметке в одном из фрагментов текста, определяются в тех разделах, которые не планировалось размечать. Работа BERT моделей из библиотек SpaCy и

DeepPavlov показала близкие результаты с небольшим преимуществом DeepPavlov. Анализ результатов показал перспективность использования указанных средств для задач нахождения именованных сущностей.

СПИСОК ЛИТЕРАТУРЫ

1. Карнаухов Н.С., Ильюхин Р.Г. Возможности технологий «Big Data» в медицине // Врач и информационные технологии. – 2019. – № 1. – С. 59–63.
2. Large-scale application of named entity recognition to biomedicine and epidemiology / S. Raza, D.J. Reji, F. Shajan, S.R. Bashir // PLOS Digital Health. – 2022. – № 1 (12). – e0000152. DOI: https://doi.org/10.1371/journal.pdig.0000152
3. Concept attribute labeling and context-aware named entity recognition in electronic health records / A. Pomares-Quimbaya, R.A. Gonzalez, O.M. Velandia, A.A. Peña, J.C. Rodríguez, A.S. Múnera, C. Labbé // International Journal of Reliable and Quality E-Healthcare (IJRQEH). – 2018. – V. 7. – Iss. 1. – P. 15. URL: http://doi.org/10.4018/IJRQEH.2018010101 (дата обращения: 21.05.2023).
4. Маслова М.А., Дмитриев А.С., Холкин Д.О. Методы распознавание именованных сущностей в русском языке // Инженерный вестник Дона. – 2021. – Т. 79. – № 7. URL: https://cyberleninka.ru/article/n/metody-raspoznavaniya-imenovannyh-suschnostey-v-russkom-yazyke (дата обращения: 21.05.2023).
5. ADPG: Biomedical entity recognition based on Automatic Dependency Parsing Graph / Yumeng Yang, Hongfei Lin, Zhihao Yang, Yujia Zhang, Di Zhao, Shuaiheng Hyai // Journal of Biomedical Informatics. – 2023. – V. 140. – P. 104317. DOI: https://doi.org/10.1016/j.jbi.2023.104317
6. MMBERT: a unified framework for biomedical named entity recognition / Lei Fu, Zuquan Weng, Jiheng Zhang, Haihe Xie, Yuqing Cao // Medical & Biological Engineering Computing. – 2023. Oct. PMID: 37833517. DOI: 10.1007/s11517-023-02934-8
7. Biomedical named entity recognition using BERT in the machine reading comprehension framework / Cong Sun, Zhihao Yang, Lei Wang, Yin Zhang, Hingfei Lin, Jian Wang // Journal of Biomedical Informatics. – 2021. – V. 118. – P. 103799. DOI: 10.1016/j.jbi.2021.103799
8. Pir Noman Ahmad, Adnan Muhammad Shah, KangYoon Lee. A review on electronic health record text-mining for biomedical name entity recognition in healthcare domain // Healthcare. – 2023. – V. 11. – № 9. – P. 1268. DOI: 10.3390/healthcare11091268
9. Meijing Li, Hao Yang, Yuxin Liu. Biomedical named entity recognition based on fusion multi-features embedding // Technol Health Care. – 2023. – № 31 (S1). – P. 111–121. DOI: 10.3233/THC-236011
10. Zhang Z., Chen A.L.P. Biomedical named entity recognition with the combined feature attention and fully-shared multi-task learning // BMC Bioinformatics. – 2022. – № 23. – Article number: 458. DOI: https://doi.org/10.1186/s12859-022-04994-3
11. Negation-based transfer learning for improving biomedical named entity recognition and relation extraction / H. Fabregat, A. Duque, J. Martinez-Romo, L. Araujo // Journal of Biomedical Informatics. – 2023. – V. 138. – Article number: 104279. PMID: 36610608. DOI: 10.1016/j.jbi.2022.104279
12. factRu. URL: https://github.com/dialogue-evaluation/factRuEval-2016/ (дата обращения: 21.05.2023).
13. Ne5. URL: https://www.labinform.ru/pub/named_entities (дата обращения: 21.05.2023).
14. Naeval – количественное сравнение систем для русскоязычного NLP. URL: https://natasha.github.io/naeval (дата обращения: 21.05.2023).
15. Spacy. Industrial-Strength Natural Language Processing in Python. URL: https://spacy.io (дата обращения: 21.05.2023).
16. DeepPavlov. URL: https://github.com/deepmipt/DeepPavlov (дата обращения: 21.05.2023).
17. SibMED Data Clinical Repository. URL: https://dataset.ssmu.ru/ (дата обращения: 21.05.2023).
18. Diederik P. Kingma, Jimmy Ba. Adam: a method for stochastic optimization. DOI: https://doi.org/10.48550/arXiv.1412.6980

Поступила: 20.10.2023
Принята: 22.11.2023

Информация об авторах

Соколовский Д.Е., аспирант отделения информационных технологий Инженерной школы информационных технологий и робототехники Национального исследовательского Томского политехнического университета.

Некрасов В.Н., врач клинически-лабораторной диагностики, Центр клинически-лабораторной диагностики Военно-медицинской академии им С.М. Кирова.

Землянский С.А., аспирант Института прикладной математики и компьютерных наук Национального исследовательского Томского государственного университета.

Аксёнов С.В., кандидат технических наук, доцент отделения информационных технологий Инженерной школы информационных технологий и робототехники Национального исследовательского Томского политехнического университета.

UDC 621.396.41
DOI 10.18799/29495407/2023/2/27

EVALUATION OF SPACY AND DEEPPAVLOV LIBRARY TOOLS FOR NAMED ENTITIES RECOGNITION FROM DESCRIPTIONS OF EXAMINATION RESULTS OF PATIENTS WITH COVID-19

Dmitry E. Sokolovsky[1],
desokolovskii@gmail.com

Vladimir N. Nekrasov[2],
nekrassov@yandex.ru

Sergey A. Zemlyansky[3],
qoelky@gmail.com

Sergey V. Axyonov[1],
axyonov@tpu.ru

[1] National Research Tomsk Polytechnic University,
30, Lenin avenue, Tomsk, 634050, Russia.

[2] S.M. Kirov Military Medical Academy,
6, lit. Zh, Academician Lebedev street, St Petersburg, 194044, Russia

[3] National Research Tomsk State University,
36, Lenin avenue, Tomsk, 634050, Russia.

Relevance. *Determined by the need to extract significant features from electronic medical records to automate the assessment of patients' condition.* ***Aim.*** *Assessing the possibility of identifying named entitie in electronic descriptions of examinations of patients with COVID-19 using the BERT model from the SpaCy and DeepPavlov libraries.* ***Methods.*** *De ep learning, statistical methods.* ***Results and conclusions.*** *The authors have carried out a fine-tuning study on BERT neural network models from the SpaCy and DeepPavlov libraries to annotate documents "Examination of patients by the attending physician" in order to highlight the following predictors of patient assessment: temperature, blood pressure, respiratory rate, heart rate and saturation. Configuration and evaluation of the effectiveness of the architectures was carried out based on the markup of 340 anonymized electronic medical records of patients with COVID-19, obtained using the SibMED Data Clinical Repository service. It is shown that setting up models on a number of about 150 labeled documents makes it possible to determine the specified predictors in such texts with accuracy (Precision) of 85–98% and completeness (Recall) of 77–98%, depending on the predictor. The quality metrics of the architectures from the selected libraries differed slightly. Iterative expansion of the training set as a result of the operation of models with subsequent additional tuning leads to an increase in the effectiveness of the models.*

Key words: *Deep learning, named entity extraction, BERT, SpaCy, DeepPavlov.*

REFERENCES

1. Karnaukhov N. S., Ilyukhin R. G. Capabilities of «Big Data» technologies in medicine. *Vrach i informatsionnye tekhnologii*, 2019, no. 1, pp. 59–63. In Rus.
2. Raza S., Reji D.J., Shajan F., Bashir S.R. Large-scale application of named entity recognition to biomedicine and epidemiology. *PLOS Digital Health*, 2022, no. 1 (12), e0000152. DOI: https://doi.org/10.1371/journal.pdig.0000152
3. Pomares-Quimbaya A., Gonzalez R.A., Velandia O.M., Peña A.A., Rodríguez J.C., Múnera A.S., Labbé C. Concept attribute labeling and context-aware named entity recognition in electronic health records. *International Journal of Reliable and Quality E-Healthcare (IJRQEH)*, 2018, vol. 7, Iss. 1, pp. 15. Available at: http://doi.org/10.4018/IJRQEH.2018010101 (accessed: 21 May 2023).
4. Maslova M.A., Dmitriev A.S., Kholkin D.O. Metody raspoznavaniya imenovannykh sushchnostey v russkom yazyke [Methods for recognizing named entities in the Russian language]. *Inzhenernyy vestnik Dona*, 2021, vol. 79, no. 7. Available at: https://cyberleninka.ru/article/n/metody-raspoznavaniya-imenovannyh-suschnostey-v-russkom-yazyke (accessed: 21 May 2023).
5. Yumeng Yang, Hongfei Lin, Zhihao Yang, Yujia Zhang, Di Zhao, Shuaiheng Hyai. ADPG: biomedical entity recognition based on Automatic Dependency Parsing Graph. *Journal of Biomedical Informatics*, 2023, vol. 140, p. 104317. DOI: https://doi.org/10.1016/j.jbi.2023.104317
6. Lei Fu, Zuquan Weng, Jiheng Zhang, Haihe Xie, Yuqing Cao. MMBERT: a unified framework for biomedical named entity recognition. *Medical & Biological Engineering Computing*, 2023. Oct. PMID: 37833517. DOI: 10.1007/s11517-023-02934-8
7. Cong Sun, Zhihao Yang, Lei Wang, Yin Zhang, Hingfei Lin, Jian Wang. Biomedical named entity recognition using BERT in the machine reading comprehension framework. *Journal of Biomedical Informatics*, 2021, vol. 118, p. 103799. DOI: 10.1016/j.jbi.2021.103799
8. Pir Noman Ahmad, Adnan Muhammad Shah, KangYoon Lee. A review on electronic health record text-mining for biomedical name entity recognition in healthcare domain. *Healthcare*, 2023, vol. 11, no. 9, p. 1268. DOI: 10.3390/healthcare11091268
9. Meijing Li, Hao Yang, Yuxin Liu. Biomedical named entity recognition based on fusion multi-features embedding. *Technol Health Care*, 2023, no. 31 (S1), pp. 111–121. DOI: 10.3233/THC-236011
10. Zhang Z., Chen A.L.P. Biomedical named entity recognition with the combined feature attention and fully-shared multi-task learning. *BMC Bioinformatics*, 2022, no. 23, article number: 458. DOI: https://doi.org/10.1186/s12859-022-04994-3
11. Fabregat H., Duque A., Martinez-Romo J., Araujo L. Negation-based transfer learning for improving biomedical Named Entity Recognition and Relation Extraction. *Journal of Biomedical Informatics*, 2023, vol. 138, article number: 104279. PMID: 36610608. DOI: 10.1016/j.jbi.2022.104279
12. *factRu*. Available at: https://github.com/dialogue-evaluation/factRuEval-2016/ (accessed: 21 May 2023).
13. *Ne5*. Available at: https://www.labinform.ru/pub/named_entities (accessed: 21 May 2023).
14. *Naeval – kolichestvennoe sravnenie sistem dlya russkoyazychnogo NLP* [Naeval – quantitative comparison of systems for Russian-

language NLP]. Available at: https://natasha.github.io/naeval (accessed: 21 May 2023).
15. *Spacy. Industrial-Strength Natural Language Processing in Python*. Available at: https://spacy.io (accessed: 21 May 2023).
16. *DeepPavlov*. Available at: https://github.com/deepmipt/DeepPavlov (accessed: 05 May 2023).
17. SibMED Data Clinical Repository. Available at: https://dataset.ssmu.ru/ (accessed: 21 May 2023).
18. Diederik P. Kingma, Jimmy Ba. *Adam: A Method for Stochastic Optimization*. DOI: https://doi.org/10.48550/arXiv.1412.6980

Received: 20.10.2023
Reviewed: 22.11.2023

Information about the authors

Dmitry E. Sokolovsky, postgraduate student, National Research Tomsk Polytechnic University.
Vladimir N. Nekrasov, Doctor of clinical and laboratory diagnostics, S.M. Kirov Military Medical Academy.
Sergey A. Zemlyansky, postgraduate student, National Research Tomsk State University.
Sergey V. Axyonov, Cand. Sc., associate professor, National Research Tomsk Polytechnic University.